Population Genetics Of Setaria Viridis, A New Model System

By: Pu Huang, Maximilian Feldman, Stephan Schroder, Bochra A. Bahri, Xianmin Diao, Hui Zhi , **Matt Estep**, Ivan Baxter, Katrien M. Devos And Elizabeth A. Kellogg

Abstract

An extensive survey of the standing genetic variation in natural populations is among the priority steps in developing a species into a model system. In recent years, green fox-tail (Setaria viridis), along with its domesticated form foxtail millet (S. italica), has rap-idly become a promising new model system for C4 grasses and bioenergy crops, due to its rapid life cycle, large amount of seed production and small diploid genome, among other characters. However, remarkably little is known about the genetic diversity in natural populations of this species. In this study, we survey the genetic diversity of a worldwide sample of more than 200 S. viridis accessions, using the genotyping-by-sequencing tech-nique. Two distinct genetic groups in S. viridis and a third group resembling S. italica were identified, with considerable admixture among the three groups. We find the genetic variation of North American S. viridis correlates with both geography and climate and is representative of the total genetic diversity in this species. This pattern may reflect several introduction/dispersal events of S. viridis into North America. We also modelled demo-graphic history and show signal of recent population decline in one subgroup. Finally, we show linkage disequilibrium decay is rapid (<45 kb) in our total sample and slow in genetic subgroups. These results together provide an in-depth understanding of the pat-tern of genetic diversity of this new model species on a broad geographic scale. They also provide key guidelines for on-going and future work including germplasm preservation, local adaptation, crossing designs and genomewide association studies.

Pu Huang, Maximilian Feldman, Stephan Schroder, Bochra A. Bahri, Xianmin Diao, Hui Zhi , **Matt Estep**, Ivan Baxter, Katrien M. Devos And Elizabeth A. Kellogg. (2014). "Population Genetics Of Setaria Viridis, A New Model System." *Molecular Ecology* Issue 23, 4912–4925 [DOI: 10.1111/mec.12907] Version Of Record Available At www.Wiley.com

Introduction

Two long-standing challenges to human society are the need to increase food production and to develop sufficient energy sources. As the global population continues to grow over the next several decades, these two challenges are becoming more important and urgent (United Nations 2013). The recent development of a biofuel industry using plant biomass as an energy source has identified a single emerging requirement for both challenges: to increase the productivity of crop species, producing grain for food crops and biomass for bioenergy crops (Brutnell *et al*. 2010).

One key effort to increase the productivity of crops is via a thorough investigation of C_4 photosynthesis (Sage

2004; Brutnell *et al.* 2010; Von Caemmerer *et al.* 2012). C_4 photosynthesis, characterized by CO_2 fixation to form a four-carbon acid, is a complex process that is found in both food (maize, sorghum and millets) and energy crops (switchgrass, sugarcane and *Miscanthus*). Compared to C_3 plants such as rice and wheat, C_4 plants maintain high photosynthetic rates in hot, dry environments, and make more efficient use of nitrogen and water (Sage 2004). However, progress in dissecting C_4 traits has been hampered by the lack of an appropriate model system (Brutnell *et al.* 2010). Existing study systems with the C_4 pathway, such as maize, sorghum, *Miscanthus* and switchgrass, have various drawbacks, including polyploidy, large genome size, long generation times and large plant size. In comparison, the recently developed model system for studying C_4 photosynthesis (and for panicoid grasses in general), green foxtail (or green millet, *Setaria viridis*) and its domesticated form foxtail millet (*S. italica*), has many advantages (Doust *et al.* 2009; Brutnell *et al.* 2010; Li & Brutnell 2011).

Green foxtail is a common annual panicoid grass distributed widely across northern and southern temperate zones around the world (Wang *et al.* 1995). Like maize, sorghum, sugar cane and *Miscanthus*, it uses the NADP-dependent malic enzyme subtype of the C_4 photosynthetic pathway. It has a diploid genome size of approximately 515 mega base pairs (Mbp) and a generation time as short as 6 weeks in greenhouse conditions. Many accessions are small plants, usually around 10-30 cm tall (field and greenhouse observations), which is convenient for both greenhouses and growth chambers. Hundreds of seeds can be easily harvested from a single inflorescence. *S. viridis* is also a naturally inbreeding species (Wang *et al.* 1995), which makes sequencing, genetic manipulations and building mapping populations relatively easy. In addition, previous studies have indicated that *S. viridis* is the closest wild relative of the domesticated foxtail millet (*S. italica,* Barton *et al.* 2009; Wang *et al.* 2010). In spite of some dramatic morphological differences, the two species are genetically very similar and interfertile (Wang *et al.* 1998; Doust *et al.* 2004; Jia *et al.* 2013a). This enables genetic tools developed in *S. italica* to be applied to *S. viridis*, including two published assembled genomes (Bennetzen *et al.* 2012; Zhang *et al.* 2012) and a *S. italica* by *S. viridis* hybrid mapping population (Wang *et al.* 1998; Doust *et al.* 2004; Bennetzen *et al.* 2012). In addition, many technical challenges have also been addressed, including making efficient crosses (Jiang *et al.* 2013) and stable transformations (Brutnell *et al.* 2010).

Despite the progress in developing *S. viridis* and *S. italica* as a new model system, information on natural diversity is very limited. Such information is essential for association mapping of desired phenotypic traits (marker discovery, marker density and control for demographic effects; Kim *et al.* 2007; Brachi *et al.* 2011; Jia *et al.* 2013a), for optimizing strategies for germplasm preservation (Sachs 2009) and for providing background knowledge for crossing designs and reverse genetic studies in general. A massive amount of data is available for the genetic variation in natural populations of already established model systems. This includes well-known examples of *Arabidopsis thaliana* (Nordborg *et al.* 2005; François *et al.* 2008; Bomblies *et al.* 2010; Platt *et al.* 2010; Cao *et al.* 2011; Long *et al.* 2013), rice (Caicedo *et al.* 2007; Huang *et al.* 2012a,b) and maize (Van Heerwaarden *et al.* 2010; Jiao *et al.* 2012; Romay *et al.* 2013). Information on genetic diversity, population structure, demographic history and linkage disequilibrium (LD) also reveal how the fundamental evolutionary forces such as genetic drift, gene flow, system of mating, recombination and natural selection interact with each other in creating and shaping the pattern of genetic variation (Hartl & Clark 2007).

Few studies have reported on natural variation in *S. viridis*. An early population genetic study using 13 isozyme markers and 168 accessions of *S. viridis* and *S. italica* from both North America and Eurasia found little differentiation between Eurasia and North America, but did identify distinct northern and southern populations in central North America, on either side of 43.5° N latitude (Wang *et al.* 1995). Jia *et al.* (2013b) used 77 microsatellite markers to survey genetic diversity of 288 *S. viridis* accessions in China. This study revealed a largely mixed geographic distribution of the genetic diversity of Chinese *S. viridis* populations and a low level of genomewide linkage disequilibrium (LD). Many studies that have focused on the domesticated *S. italica* (e.g. D' Ennequin *et al.* 2000; Wang *et al.* 2010, 2012; Jia *et al.* 2013a) have included a few samples of *S. viridis*, but the small numbers of plants included do not permit detailed population genetic analysis. While these studies have shaped our understanding of *S. viridis* diversity, many questions remain to be answered. Population genetic structure of *S. viridis* needs to be re-evaluated at a genomewide scale and with better geographic coverage. Also, the extent of LD needs to be quantified across the genome and at high resolution (down to approximately 1 kb). This information is critical for crossing design, association studies and germplasm preservation strategies (Kim *et al.* 2007; Sachs 2009; Brachi *et al.* 2011; Jia *et al.* 2013a). In addition, many population genetic topics interesting in their own right, including biogeography, demographic history and adaptation to local environments, could be examined in much more detail given data sets with good genomic

and geographic coverage (e.g. Nordborg *et al.* 2005; Kim *et al.* 2007; Van Heerwaarden *et al.* 2010; Huang *et al.* 2012b).

In this study, we report on the genetic diversity of 217 *S. viridis* and *S. italica* accessions collected worldwide but with a focus on the temperate zone of North America, using the genotyping-by-sequencing (GBS) approach (Elshire *et al.* 2011) and the assembled *S. italica* genome (Bennetzen *et al.* 2012) as the reference. The central goal is to examine the extent and pattern of genetic variation in natural populations of *S. viridis* at the genomewide scale. Specifically, we aim to answer four questions in our study system: (i) what is the population genetic structure; (ii) how is genetic variation related to geography and climate; (iii) what is the demographic history of different genetic groups; and (iv) what is the pattern of linkage disequilibrium. We further discuss the significance of our findings in terms of developing *S. viridis* as a model system and how they may influence future work.

Materials and methods

Germplasm and sequencing

A total of 252 *S. viridis* and 21 *S. italica* accessions were obtained from various sources. Majority of the *S. viridis* accessions were collected in North America, and the remaining accessions originated mainly from Eurasia (Table S1, Supporting information). Plants were grown in the glasshouse at the University of Georgia, Athens, GA, and at the Donald Danforth Plant Science Center, St Louis, MO. Due to the continuous morphological variation in natural populations of *S. viridis* and its close relatives (e.g. *S. faberi*), morphological characters of these individuals were carefully examined at the time of both collecting and growing. High molecular weight DNA of each individual was extracted from young leaves using a CTAB protocol. Genotyping by sequencing (GBS) was performed largely using the protocols developed by Elshire *et al.* (2011) and Poland *et al.* (2012). Briefly, DNA was double-digested with the restriction enzymes *Pst*I and *Ms*pI, a barcoded adaptor was added on the *Pst*I site and a common Y adaptor was added to the *Ms*pI site. Following ligation, fragments <300 bp were removed from individual samples using 0.7 volumes of Sera-Mag™ Magnetic SpeedBeads prepared according to Rohland & Reich 2012). Barcoded samples were pooled (96-plex or 124-plex) and used for library construction. The PCR extension time during the library preparation step was limited to 15 s to reduce amplification of larger DNA fragments. The 96-plex and 124-plex pools were sequenced (100-bp paired-end reads) on one lane of an Illumina Hi-Seq using the high-output run mode or 2 lanes (on-board clustering) of a HiSeq using the rapid run mode, respectively.

Data processing and handling

Setaria viridis and *S. italica* accessions were grouped together for most data processing and data analysis unless otherwise specified; because *S. italica* is known to be domesticated from *S. viridis* about 10 000 years ago, most of its genome is highly similar to that of *S. viridis*, and the majority of its genetic variation is expected to be a subset of *S. viridis* (Barton *et al.* 2009; Wang *et al.* 2010). Paired-end raw fastq read pairs of all accessions were split using an in-house script based on their barcodes into separate fastq files. All reads were quality-trimmed at the 3′ end based on average Phred scale quality scores <20 using a 5-bp sliding window and then again with a 1-bp sliding window. We did not use the existing GBS pipeline TASSEL (Bradbury *et al.* 2007), because TASSEL handles only single-end reads, whereas our data were paired-end reads. The quality-trimmed sequences were aligned to the nonmasked reference sequence of *S. italica* (JGI 2.0.16, ftp://ftp.ensemblgenomes.org/pub/plants/release-16/fasta/setaria_italica/dna/) using bowtie2 (Langmead & Salzberg 2012). Only mapped reads with a Phred scale quality score <20 were retained for downstream analysis. Local realignment was performed using GATK (McKenna *et al.* 2010). Single nucleotide polymorphisms (SNPs) were called using UnifiedGenotyper of GATK for all accessions that passed the quality filter together.

We reasoned that false positive SNPs (i.e. calling a SNP that was not present) would distort the signal in the data, whereas false negative SNPs (i.e. missing a SNP) would simply reduce the sample size. To minimize the number of false positives, a series of conservative filters was applied to the raw SNP calls to ensure high-quality data: only those SNPs (i) with missing data <10% and low-quality data (see below) <15%, (ii) with only one alternative allele, (iii) mapped to one of the nine chromosomal scaffolds, (iv) with Phred scale mapping quality <35 and (v) total read depth >100 were retained. Individual genotypes were regarded as low quality if individual read depth was smaller than 8, or individual Phred scale genotype call quality was <20; low-quality calls were subject to the 15% filtering mentioned above. Finally, any individual with more than 2% missing genotypes was also excluded from analysis.

The inbreeding coefficient F_{IS} was estimated for each SNP, and the selfing rate was estimated using average $s = 2F_{IS}/(1 + F_{IS})$ for all SNPs (Hartl & Clark 2007) with minor allele frequency (MAF) <0.01. Phasing and missing data imputation was then carried out using BEAGLE 3.3.2 (Browning & Browning 2007) using the

genotype likelihoods. As *S. viridis* is known to be mainly self-pollinating (Wang *et al.* 1995) (also see results of this study), SNPs with inbreeding coefficient $F_{IS} < 0.5$ are likely to contain sequencing and/or mapping errors and were discarded (12.4% of the total SNPs discarded). Genotypes of the reference *S. italica* accession Yugu1 reference genome were added to the data set at the last step. Finally, 217 individuals with 39 416 high-quality SNPs were obtained for downstream data analysis.

Genome and marker properties

All SNPs were assigned different categories (nonsynonymous, synonymous, intron and intergenic) using snpEff 3.2 (Cingolani *et al.* 2012). The transition to transversion ratio was estimated for all SNPs together and SNPs in each category separately. Because methylation-sensitive restriction enzymes were used in library preparation, GBS tags were expected to avoid heavily methylated repetitive regions, and the CpG effect was expected to be weak compared to whole-genome sequencing. The 3′ flanking base of each SNP was examined to illustrate this point. Local SNP densities were estimated using a sliding window of 1 Mb with 0.2 Mb increments on each of the nine chromosomes. The minor allele frequency (MAF) spectrum of each SNP category was also calculated and compared.

Population genetic structure

Genetic clusters in *S. viridis* and *S. italica* were identified using STRUCTURE (Falush *et al.* 2007). This analysis was performed six times per cluster number value (K) from 1 to 9 with random starting points. We first set both the burn-in and actual run length to 10 000 steps to determine the deltaK statistic (Evanno *et al.* 2005). This method showed that K = 2 and 3 are the most reasonable cluster numbers for our data (Fig. S1, Supporting information). We then performed two long runs of both burn-in and actual run length of 40 000 steps for K = 2 and 3. Nonadmixed individuals in each genetic group were determined using Q-matrix assignment of above 0.9. Principal component analysis (PCA) was performed using the smartpca program of the EIGENSOFT package (Patterson *et al.* 2006). Outlier iterations were not performed to make the result compatible with other analyses. The first three principal components (PCs) were examined. We used software NJTREE (http://treesoft.sourceforge.net/treebest.shtml) to construct a neighbour-joining tree, using SNPs with minor allele frequency more than 0.1. Branch support was calculated based on 10 000 bootstrap iterations. The results from the three different approaches were then summarized and compared. Analysis of molecular variance (AMOVA) was performed using Arlequin 3.5 (Excoffier & Lischer 2010) on nonadmixed individuals of each group (>0.9 assignment in Q matrix). Both overall and pairwise F_{ST} were calculated.

Correlation with geography and climate

To examine the relationship between genetic variation and geographic locality and to elucidate the possible origin of North American *S. viridis*, all *S. viridis* samples with available coordinates were plotted on a map and coloured according to their genetic group assignment using qGIS 4.0 (www.qgis.org). We further correlated geographic distances (great circle distances, distance between two points corrected by the curvature of the earth's surface) and genetic distances (neighbour-joining) among *S. viridis* accessions from North America. The Eurasian and South American accessions were not included because cross-ocean geographical distances are much less biologically meaningful. We also correlated climatic and genetic distances. Values of 19 bioclimatic variables (Table S2, Supporting information) at known collecting localities were extracted from the corresponding climatic layers (www.worldclim.org, Hijmans *et al.* 2005) using qGIS 4.0. Each climatic variable was standardized by subtracting the mean value and then divided by the standard deviation across the sample. The distance matrix of the standardized climatic variables was then correlated with the genetic distance matrix. An association study of the 19 bioclimatic variables was also performed using the mixed linear model method from TASSEL 4.0 (Bradbury *et al.* 2007) to detect potential genomic regions responding to climatic conditions. The Q matrix obtained from the STRUCTURE analysis was used as covariate.

Demographic history

As the population structure analyses identified two very distinct genetic groups in *S. viridis* and a *S. italica*-like Group 3, part of which has a known complicated domestication history, we restrict our demographic history inference to the two *S. viridis* groups and using only the nonadmixed individuals. We used software δaδi (Gutenkunst *et al.* 2009) to compare different demographic models by fitting their predicted MAF spectra (the lowest frequency bins are masked) to the observed data. We examined three different population demographic models: (i) the standard neutral model (SNM), assuming a constant population size (ancestral population size, N_a) throughout time; (ii) simple growth model, assuming population size has grown at a constant rate since time T in the past, to yield a present

population size of N_f; and (iii) bottleneck + growth model, assuming population size experienced a sudden change to N_b at time T in the past, followed by a constant growth rate until present, yielding a present population size of N_f. The simple growth model is a special case of the bottleneck + growth model (when $N_a = N_b$), and the SNM is a special case of the simple growth model (when $N_a = N_f$). Parameters in these three models were optimized to fit the observed data in each genetic group. To identify the best fit demographic model, likelihood ratio tests were performed between SNM vs. optimized simple growth model (degree of freedom=2-0=2), and optimized simple growth model vs. optimized bottleneck + growth model (degree of freedom=3-2=1), in Group 1 and Group 2 separately.

Linkage disequilibrium

Linkage disequilibrium (LD) is measured between pairs of SNPs with a MAF of more than 0.05 on the same chromosome using r^2. Decay of LD was estimated by fitting the theoretical LD decay formula (Hill & Weir 1988; Remington *et al.* 2001) using the least-squares criterion. We also examined LD using predefined distance intervals of physical distances (0, 5, 10, 15, 20, 50, 100, 150, 200, 250, 300, 400, 500, 750 and 1000 kb), using 25%, 50% and 75% quantile of r^2. Both analyses were performed using all samples together and nonadmixed individuals of each genetic group (<0.9 assignment).

Results

Genome and marker properties

The GBS tags on average covered about 3.8% of the total reference Yugu1 genomic sequence. The majority of the GBS tags were distributed on the distal portions of the chromosome arms, and far fewer were located in regions near the centromeres (Fig. 1A). The total number of SNPs after all filters were applied is 39 416. The

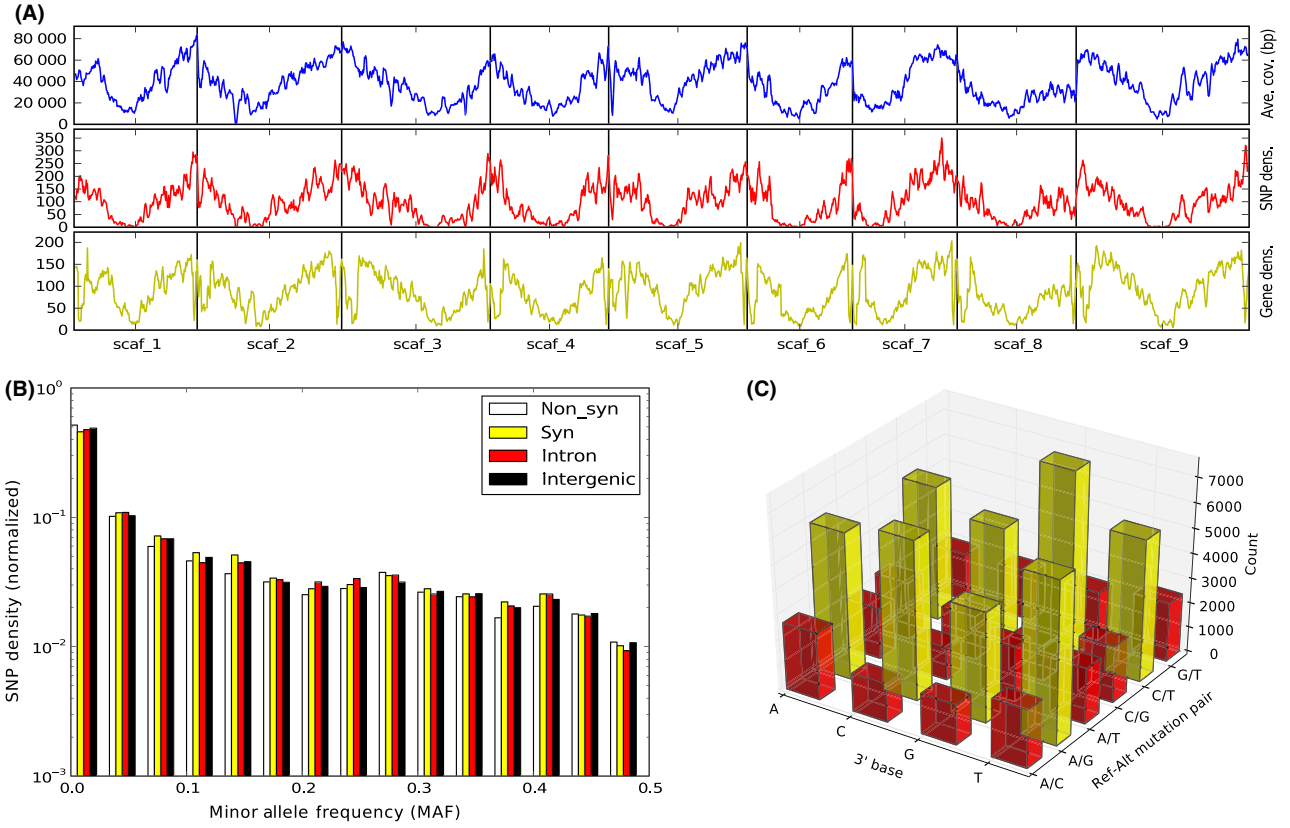

Fig. 1 Marker properties of *S. viridis* (A) Distribution of average (per individual) sequence coverage of GBS tags, density of SNPs and density of genes based on the *S. italica* reference genome (Yugu1) showing nine chromosomal scaffolds. The sliding window size is 1 Mb with increments of 0.2 Mb. (B) Minor allele frequency spectra of the total sample. SNPs are categorized based on their functional groups. (C) Prevalence of SNPs based on reference-alternative nucleotide pair and 3′ flanking nucleotide. All SNPs are considered on both forward and reverse complementary direction, and forward and reverse mutational directions are combined here as there is no valid outgroup. SNP counts in each category are also shown in Table S3 (Supporting information).

average selfing rate is 0.9610 (SD 0.0426, before filtering of $F_{IS} > 0.5$). Most of these SNPs (approximately 77.0%) are in noncoding intergenic and intron regions (21 328 and 9003, respectively). In coding regions, nonsynonymous and synonymous SNPs occur in a ratio of approximately 1.05 (3408 and 3244, respectively). There are also many more SNPs in 3′ UTRs than in 5′ UTRs (2273 and 412, respectively). Distribution of the SNPs largely corresponds to the distribution of GBS tags, with high density on chromosome arms and low density near centromeres (Fig. 1A).

Using the minor allele frequency (MAF) spectrum of intergenic SNPs as a neutral reference, MAF spectra of the other three categories (nonsynonymous, synonymous and intron sites) are all similar to the neutral reference, with no obvious excess of low-frequency SNPs (Fig. 1B). The genomewide transition to transversion ratio (observed events) is 1.45 (ranging from 1.33 to 1.51 on each chromosome, Table S2, Supporting information), and intergenic regions tend to have lower transition to transversion ratios (1.32) than coding and intron regions (1.96 and 1.45; Table S2, Supporting information). There is also a slightly elevated preference of CpG-TpG mutations compared to other transitions (Fig. 1C; Table S3, Supporting information).

Population structure

DeltaK values from the STRUCTURE analysis are highest for two clusters and moderate for three clusters (Fig. S1, Supporting information). The latter value is biologically more meaningful as it distinguishes the domesticated *S. italica* from *S. viridis* (Fig. 2A; Table S1, Supporting information). The three genetic groups thus include two major groups in *S. viridis*, and a third group consisting of *S. italica* and *S. italica*-like individu-

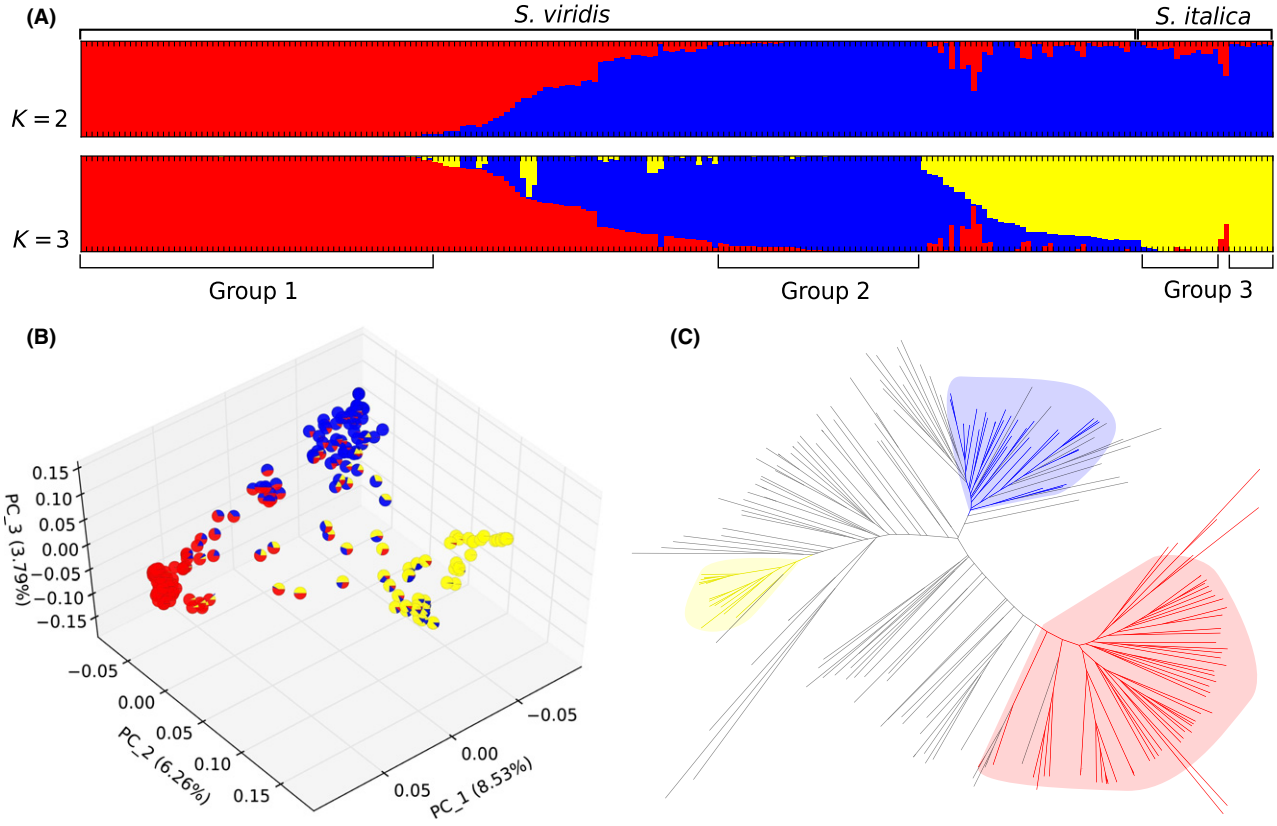

Fig. 2 Population structure of *S. viridis* and *S. italica*. (A) STRUCTURE result. The top panel is a result of K = 2, and bottom panel is a result of K = 3 (highest likelihood result among replicates). Group 1, Group 2 and *S. italica*-like Group 3 are colour-coded in red, blue and yellow, respectively. Accessions marked by group name are considered nonadmixed (more than 0.9 assignment to one group in STRUCTURE analysis with K = 3). (B) Result of principal component analysis. Each small pie chart corresponds to an individual, and the different colour slices correspond to the group assignment matrix (Q matrix) of the STRUCTURE result for K = 3. Numbers in parentheses beside each axis denote the amount of variance explained by that axis. (C) Neighbour-joining tree. The red-, blue- and yellow-coloured backgrounds approximately correspond to Group 1, Group 2 and Group 3. Coloured branches represent the nonadmixed individuals within each corresponding group. Grey branches represent admixed individuals (<0.9 assignment assigned to any group in STRUCTURE analysis with K = 3). Numbers on branches represent bootstrap support.

als (referred to as Group 1, 2 and 3 for convenience; Groups 2 and 3 generally merge into a single group when K = 2). There are many admixed accessions between the two groups of *S. viridis*, and between Group 2 of *S. viridis* and the *S. italica*-like Group 3. Admixed individuals between Group 1 and Group 3 are less common. Smaller subgroups within each genetic group are observed when large cluster number is applied, yet the distinction among the three major groups largely remains (Fig. S2, Supporting information).

The results of PCA are largely similar to those from STRUCTURE when three clusters are taken: the three genetic groups are distinct in the principal component space of three major PCs, while individuals showing a high level of admixture tend to reside between the dot clouds of their corresponding source groups (Fig. 2B). The three major PCs from the PCA explain 18.58% (8.53%, 6.26% and 3.79%, respectively) of the total variance.

The neighbour-joining tree shows that each of the three groups corresponds to a distinct branch on the tree, although the nonadmixed individuals cluster with some admixed individuals (Fig. 2C, Fig. S3, Supporting information). These admixed individuals tend to be genetically largely similar to the group corresponding to their branch, yet with a small amount of introgression from other groups. Admixed individuals with approximately equal amounts of two types tend to root in between the three branches (Fig. 2C, Fig. S3, Supporting information).

Defining 'nonadmixed individuals' as those with at least 90% of their markers from a single group gives 66 individuals in Group 1, 37 individuals in Group 2 and 22 individuals in Group 3 (Table S1, Supporting information). Both the standard and locus-by-locus AMOVA found an average F_{ST} of 0.491 across all nonadmixed individuals. The pairwise F_{ST} is 0.499 between Group 1 and Group 2, 0.447 between Group 2 and Group 3 and 0.539 between Group 1 and Group 3.

Correlation with geography and climate

The geographic pattern of genetic variation is complex. All three groups have individuals in both North America and Eurasia. In central North America, the two genetic groups of *S. viridis* are differentiated latitudinally, with Group 1 largely distributed in the northern USA and Canada, while Group 2 is mostly distributed in the central USA; however, this north–south division breaks down in the Pacific northwest (Fig. 3). Some admixed individuals between Group 3 and Group 2 are found in the eastern USA, between the range of Group 1 and Group 2. In Eurasia, accessions from Europe mostly belong to Group 1 and accessions from west to Central Asia mainly belong to Group 2, although with some admixture. *S. viridis* accessions from China are either the *S. italica*-like Group 3 or admixtures between Group 2 and Group 3 (Fig. 3). Geographic distances are significantly correlated with genetic distances among all North American accessions ($r = 0.311$, $P < 0.0001$, Fig. 4A).

Genetic distances also strongly correlate with environmental distances based on 19 bioclimatic variables (Table S1, Supporting information; Hijmans *et al.* 2005) from the worldclim database ($r = 0.373$, $P < 0.0001$, Fig. 4B). This correlation between climatic and genetic distances is stronger than the correlation between geographic distances and genetic distances. To identify genomic regions that might harbour genes for adaptation to specific climatic conditions, we further undertook a genomewide association analysis of individual

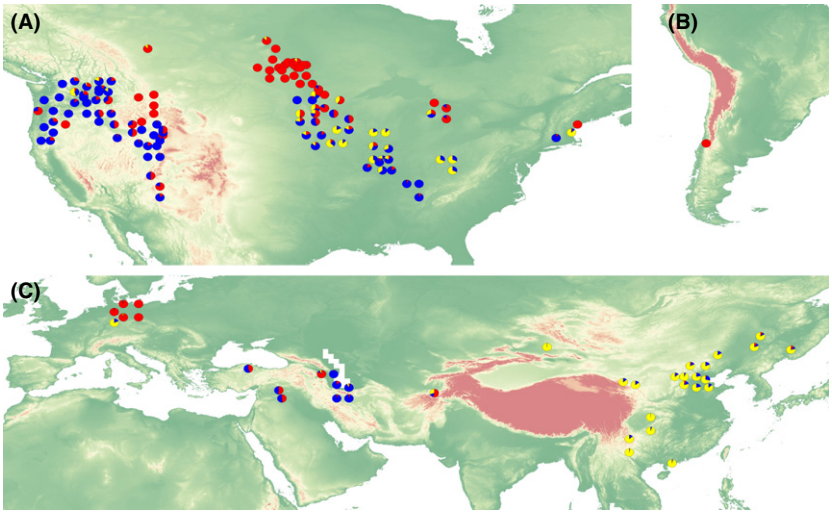

Fig. 3 Geographic distribution of *S. viridis*. All accessions with available coordinates in our study (197 individuals) are shown on the map. Each small pie chart corresponds to an individual, and the different colour slices correspond to the group assignment matrix (Q matrix) of the STRUCTURE result with K = 3. (A) North America. (B) South America. (C) Eurasia.

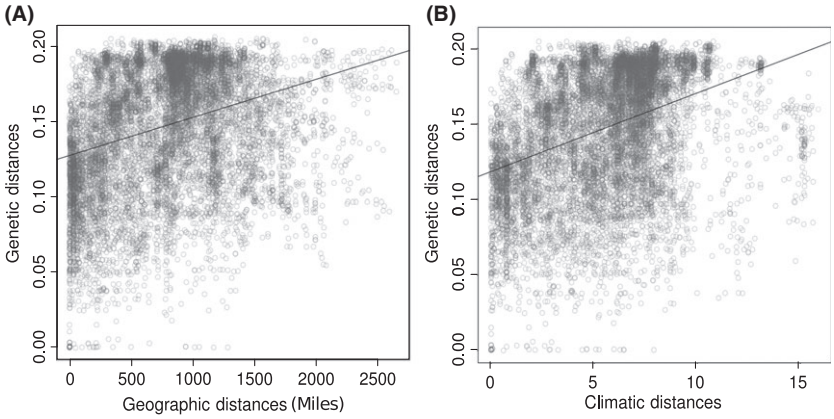

Fig. 4 Correlation analysis between genetic distances, geographic distances and climatic distances. The correlation analysis is based on the North American *S. viridis* sample. The black line represents the regression line of the total sample. (A) Correlation between genetic and geographic distances. (B) Correlation between genetic and climatic distances.

variables (Table S1, Supporting information). Association mapping of 18 of the 19 bioclimatic variables showed no significant associations after Bonferonni correction. Precipitation of warmest quarter (Bio18) did show a significant association with several regions of the genome. However, the Q–Q plot showed a very prominent deviation from the expected diagonal (Fig. S4, Supporting information), indicating a strong non-normal distribution of residuals and a very high false positive rate.

Demographic history

The observed MAF spectra of the two *S. viridis* groups are quite different from each other, and also from the neutral expectation: the MAF of Group 1 is biased heavily towards moderate frequency alleles, and the MAF of Group 2 largely shows a slight bias towards the low frequency (Fig. S5, Supporting information). In both Group 1 and Group 2, simple growth models are significantly better than the SNM and are not significantly worse than the bottleneck + growth model based on likelihood ratio test (Table 1). Specifically, in Group 1, a strong population decline is suggested by the model: the effective population size of this group has reduced to approximately 1.2% of its ancestral population size (N_a) in 0.2 N_a generations. In contrast, Group 2 has grown moderately into approximately 3.3 times of its N_a, in a very prolonged time span of 9.0 N_a (Table 1).

Linkage disequilibrium (LD)

Using the least-squares fit to a theoretical LD formula showed that local LD decay is rapid when all samples are considered together (r^2 decays to half value of 0.25 at approximately 45 kb; Fig. 5). In contrast, local LD decay is slow in nonadmixed individuals of Group 1 (half decay of r^2 at approximately 190 kb) and Group 3 (half decay of r^2 at approximately 130 kb) and extremely slow in Group 2 (half decay of r^2 at approximately 800 kb; Fig. 5). Using the quantiles of predefined distance intervals, the observations still hold that LD of the total sample decays faster than in the genetic subgroups, and Group 2 has a very slow LD decay rate.

Table 1 Comparison of demographic models

		Simple growth				Bottleneck + growth				
Group	SNM ln(L)	ln(L)	*P*	N_f	*T*	ln(L)	*P*	N_b	N_f	*T*
1	−371.8	−40.54	0.000*	0.012	0.207	−40.60	1.000	0.464	0.027	0.237
2	−66.47	−40.55	0.000*	3.32	9.00	−38.73	0.057	5.823	0.845	0.471

SNM: standard neutral model.
ln(L): log likelihood of the model.
P: chi-square probability of likelihood ratio test (simple growth vs. SNM comparison, or bottleneck + growth vs. simple growth comparison).
N_f: Present effective population size (in unit of ancestral population size, N_a).
T: Time of 'bottleneck' (in unit of N_a).
N_b: Effective population size at 'bottleneck' (in unit of N_a).
*The more comprehensive model is significantly ($P < 0.01$) better than the simpler model.

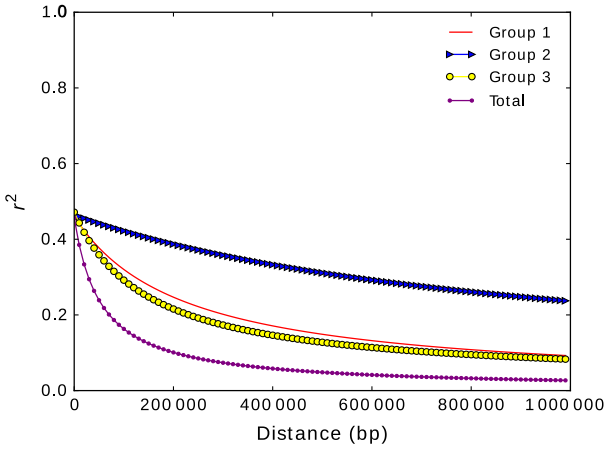

Fig. 5 Decay of linkage disequilibrium (LD). The best fitting curve of decay of linkage disequilibrium in total sample, non-admixed individuals of Group 1, Group 2 and Group 3.

The rate of estimated LD decay generally tends to be a bit faster than found by formula fitting, especially for the total sample (median r^2 drop to 0.1 at approximately 10 kb; Fig. S6, Supporting information).

When the entire genome is considered for the total sample, Group 1 alone, or Group 3 alone, blocks of high LD are fairly small (Fig. S7, Supporting information). In contrast, strikingly large LD blocks are observed in Group 2 on multiple chromosomes; these blocks are seen not only in pericentromeric regions, but also on chromosome arms (Fig. S7, Supporting information). Because such a pattern could be created by comparing two distinct uniform populations, we examined which plants shared the same haplotypes within the large LD blocks. Close examination of two blocks on Chromosome II and one on Chromosome IV revealed long haplotype sharing by different individuals in each LD block. One block (from approximately 5-26 Mb on Chromosome II) is due to a split between eastern North American vs. western North American and Eurasian plants. Another block (from approximately 42-47 Mb on Chromosome II) reflects that some western North American plants (ten accessions) share a unique long haplotype distinct from other plants. The third block (from approximately 7-32 Mb on Chromosome IV) is due to two accessions from the northwest USA that share a long haplotype distinct from other plants.

Discussion

Genome and marker properties

Our estimate of the selfing rate using SNP data is 0.9610, although the actual rate may be slightly higher due to a small number of remaining sequencing/mapping errors. This result agrees with previous findings that *S. viridis* is a highly self-pollinating species (Wang *et al.* 1995). The selfing rate is comparable to that of other inbreeding model systems such as *Arabidopsis thaliana* (approximately 97%, Platt *et al.* 2010) and rice (>95%, Oka & Morishima 1967). The high homozygosity (as a result of selfing) in *S. viridis* and *S. italica* provide many technical advantages for this study system, including the ease of sequencing, genome assembly and testing genetic by environment interactions.

The GBS tags in this study are concentrated near the distal ends of the chromosomes and closely track the gene density (Fig. 1A). This demonstrates the effectiveness of using the methylation-sensitive restriction enzyme *Pst*I to enrich for SNPs in genic regions. The high similarity between the GBS and SNP distribution profiles (Fig. 1A) indicates the absence of any large regions with little genetic diversity in the *S. viridis* germplasm analysed. The genomewide transition to transversion rate ratio (1.45) is much lower than similar estimates in other species such as *Arabidopsis thaliana* (2.4, Ossowski *et al.* 2009) and maize (2.5, Jiao *et al.* 2012). This is most likely due to the fact that GBS tags are under-represented in repetitive DNA where levels of CpG methylation and hence the potential for deamination of 5-methyl-cytosine to thymine are high, leading to high transition rates (e.g. Morton *et al.* 2005). This also explains why the CpG effect is not very strong in our data (Fig. 1C, Table S2, Supporting information).

The similarity of the overall MAF spectrum of nonsynonymous SNPs to a 'neutral' MAF spectrum (synonymous, intron and intergenic SNPs, Fig. 1B) is surprising; usually nonsynonymous mutations are subject to purifying selection, so their MAF spectrum should be biased towards low-frequency SNPs (e.g. Nordborg *et al.* 2005; Kim *et al.* 2007). As we are comparing different SNP types within the same sample, demographic factors that have genomewide effects, such as population structure, also cannot account for the observation. One possible cause for this observation is that *S. italica* gene models may not be precise enough for SNP annotations in *S. viridis*. On the other hand, it could also be that for majority of the genes, purifying selection is not strong enough (compared to genetic drift) to keep new nonsynonymous alleles at low frequencies.

Population structure and geography

All three analyses, STRUCTURE, PCA and neighbour-joining, uncover the same pattern of two very distinct genetic groups in the *S. viridis* sample, one *S. italica*-like group and a number of individuals representing mixtures (Fig. 2). On one hand, the F_{ST} values of the nonadmixed individuals from these three groups are

exceedingly high (0.491 by AMOVA) compared to those observed in large scale samples in other species such as *Oryza rufipogon* (approximately 0.1, Huang *et al.* 2012b) and *A. thaliana* (approximately 0.2, Long *et al.* 2013), indicating substantial differentiation among subgroups. On the other hand, a large number of accessions with various levels of mixture among these three groups are observed. This explains why 'monophyly' of nonadmixed individuals is not observed in the neighbour-joining tree (Fig. 2C). The admixed individuals between Group 1 and Group 2 do not show elevated heterozygosity (Table S1, Supporting information; Fig. S8, Supporting information), meaning they are likely to be results of ancient cross-pollination events followed by many generations of inbreeding. Some of the Group 3 accessions and the admixed individuals between Group 2 and Group 3, which are mainly found in China, do show elevated heterozygosity (Table S1; Fig. S8, Supporting information), suggesting more recent or ongoing cross-pollination may occur in that area. In summary, in spite of being a self-pollinating species, natural populations of *S. viridis* show clear population subdivision and have a large amount of genetic diversity resulting from historical mixtures among subgroups. This pattern is similar to what has been observed in the model plant, *A. thaliana*, which is also a self-pollinating weed with wide geographic distribution (Nordborg *et al.* 2005; Kim *et al.* 2007; Platt *et al.* 2010).

The genetic diversity in *S. viridis* also shows a clear geographic pattern. In central North America, the two *S. viridis* groups (Group 1 and Group 2) show a north–south distribution pattern, with a few admixed individuals between Group 1 and Group 2 occurring in between. This is consistent with the previously identified 43.5° N genetic dividing line of North American *S. viridis* (Wang *et al.* 1995), but with more extensive genomic and geographical coverage. The north–south distinction breaks down, however, in the Pacific Northwest, which has mainly Group 2 accessions. This indicates that the distinction between the two subpopulations may not simply reflect adaptation to day length variation in different latitudes. A few accessions from parts of the central and southern USA also show a *S. italica*-like Group 3 component (Fig. 3A). This is not surprising, given that *S. italica* has been introduced and cultivated broadly in USA for hay and bird seed (Baltensperger 2002). Such accessions could be de-domesticated from *S. italica*, hybrids between introduced *S. italica* and local *S. viridis*, or seed contaminants from *S. viridis* during importing.

Our samples from Europe mostly resemble Group 1, while samples from Central Asia mainly resemble Group 2. This differs from the previous study by Wang *et al.* (1995), who suggest no clear geographic pattern in Eurasia. As our study has better genomic coverage yet with fewer randomly distributed samples from Eurasia, this discrepancy can be solved by incorporating more accessions from Eurasia in the future. The Chinese *S. viridis* accessions, on the other hand, mostly tend to belong to the *S. italica*-like Group 3 (Fig. 2A; Fig. 3; Table S1, Supporting information); the general uniformity of this group may help explain an earlier finding that *S. viridis* in China lacks strong population structure (Jia *et al.* 2013b) and provides an interesting contrast to the clear population structure of North American *S. viridis*. This result is also in accordance with many previous studies which pointed out that the domestication centre of *S. italica* is northern China (Barton *et al.* 2009; Wang *et al.* 2010, 2012).

This pattern of geographic distribution of genetic variation in *S. viridis* could result from multiple introductions into North America from distinct gene pools in Eurasia, China, Central Asia and Europe. We find no evidence that genetic diversity in North America is simply a subset of that in Eurasia, but rather it is a nice representation of the total genetic diversity in this species. Thus, our germplasm collection of *S. viridis* accessions only from North America could be more representative of the whole species than its geographic coverage suggests. It appears that *S. viridis* is unlike other introduced weedy species such as *A. thaliana* which experienced a single recent introduction into North America (likely human mediated) followed by expansion (Jørgensen & Mauricio 2004; Nordborg *et al.* 2005; Platt *et al.* 2010). However, to fully prove this hypothesis, a much more thorough collection from the Old World is required, and this is one possible direction for future work.

Local adaptation to climate

Genetic distance correlates with geography ($r = 0.311$) and even more strongly with climate ($r = 0.373$) in North American *S. viridis*. Common hypotheses for such correlations include the balance between genetic drift and gene flow (most likely to be seed dispersal in *S. viridis*), historical factors (e.g. multiple introductions) and/or adaptation to local environment. It would be ideal to identify potential loci responsible for local adaptation (Fournier-Level *et al.* 2011; Long *et al.* 2013; Savolainen *et al.* 2013). However, we failed to detect significant associations of genotypic variation with bioclimatic variables. The strong population genetic structure, which leads to large variance in neutral F_{st}, may overwhelm detectable signals of local adaptation. The bioclimatic variables correlate with genomewide neutral variation, but this correlation largely disappears after controlling for population structure and kinship

(Fig. S4, Supporting information; Fournier-Level *et al.* 2011). In addition, climatic variables tend to be spatially autocorrelated, and as our samples are not completely evenly distributed, some of the bioclimatic variables (e.g. Bio 18) strongly deviate from a normal distribution, causing strong bias in the Q–Q plot and false positive signals.

Demographic history

Fitting our data to a demographic model shows a strong population decline in Group 1 and prolonged population growth in Group 2. A previous study suggested that the *S. viridis* populations in North America might be a result of a very recent post-Columbian human-mediated introduction (Dekker 2003). This should lead to signals of extremely rapid population growth and range expansion following the introduction, for example little population structure, no strong spatial pattern of genetic variation, strong excess of low-frequency SNPs and long shared haplotypes. For example, such signals are clearly seen in *A. thaliana*, which is also a selfing weedy species and believed to have a very recent introduction history to North America (Nordborg *et al.* 2005). However, the signals of population decline in Group 1, slow population growth in Group 2, together with the strong geographic pattern of different genetic groups in North America, and the likely multiple sources of different genetic groups, seem to not be fully compatible with a simple post-Colombian introduction. The actual history could be a more complex scenario, for example, one possibility is that Group 1 may be a more ancient introduction to North America that has experienced glacial cycles, which would explain the population decline. Group 2, on the other hand, could potentially be a more recent introduction, as long shared haplotypes are observed in many individuals of this group (Fig. S7, Supporting information).

Linkage disequilibrium and its implication for association studies

Decay of LD (using SNPs with MAF larger than 0.05) in the total population (r^2 reaches 0.25 at approximately 45 kb, Fig. 5) is much faster than in the three subgroups considered individually (in all three groups r^2 did not reach 0.25 within 100 kb; Fig. 5). The large number of admixed individuals seems to account for the rapid LD decay in the overall sample, a pattern similar to that observed in *A. thaliana*. Presumably their global distribution and large effective population size enables even a small amount of outcrossing to break down LD effectively in the total population, yet retain high levels of LD in local populations (Kim *et al.* 2007; Cao *et al.* 2011). The actual rate of LD decay in the total sample is sensitive to the analysis method that is used. Nonetheless, a conservative estimate is that the decay of LD in the total sample is within 45 kb. This result is also generally similar to that observed in *A. thaliana* (approximately 20 kb, Nordborg *et al.* 2005; Kim *et al.* 2007).

In genetic subgroups, LD decay in Group 1 and the *S. italica*-like Group 3 (r^2 reaches 0.25 at approximately 160 k) is slightly slower but comparable to the estimates from a much larger *S. italica* sample (r^2 reaches 0.25 at approximately 100k, Jia *et al.* 2013a), and also to the LD decay in cultivated rice (approximately 123 kb for *indica* and approximately 167 kb for *japonica*, Huang *et al.* 2012a). On the other hand, the slow LD decay observed in Group 2 (Fig. 5) reflects the large high-LD blocks caused by long haplotypes (Fig. S7, Supporting information). At least some of these blocks (e.g. 42.0-47.4 Mb on Chromosome II) lie in regions that otherwise have high physical recombination rates (Bennetzen *et al.* 2012). The same LD blocks are not seen, or are much smaller, in other subgroups. Thus, the LD blocks observed in Group 2 do not reflect simple physical linkage. Also, as different haplotypes in various LD blocks were not shared by the same groups of individuals, fine scale population structure may not explain this observation either. The pattern could be an artefact of geographical sampling, as our sample of Group 2 is divided between eastern and western regions (Fig. 5). Filling the 'gap' between the eastern and western USA might be able to find recombinants that break up the large LD blocks.

Population structure is known to bias genomewide association studies (GWAS) and can cause false positive associations (Brachi *et al.* 2011; Korte & Farlow 2013). Given the observed strong population structure in *S. viridis*, it might be difficult to do GWAS on the total sample. Solutions to the problem are either to do GWAS in each subgroup separately (resampling method, Brachi *et al.* 2011), or to restructure the mapping population to use only a few 'representative' individuals from each subgroup (restructure method, Brachi *et al.* 2011). We currently do not have enough individuals to sample within either of the subgroups. In addition, the large, nonrandomly distributed LD blocks in Group 2 (Fig. S7, Supporting information) preclude high-resolution GWAS in this sample. The admixed individuals could potentially be useful for the restructure strategy, because LD decay is expected to be relatively fast in this group (less then approximately 45 kb). The SNP density of the data set used in the current analysis may be insufficient (one SNP per approximately 17 kb region on average), considering many SNPs have low minor allele frequencies. However, GBS tags and SNPs are largely distributed in gene-dense regions (Fig. 1A). It is possible that, by relaxing some of

the filtering criteria (e.g. accept 20% missing rate and read coverage of 5×), and using more rigorous imputation methods (Marchini & Howie 2010), we could potentially obtain more than 200 000 SNPs. This would allow a restructured version of our current data set to be applied to trait–marker associations using GWAS, once phenotypic data become available for this diverse set of *Setaria* accessions.

Acknowledgement

We thank D. Vela, K. Waselkov, J. Thompson, P. Sweeney, C. Roché, J. Penagos, M. Weigend, H. Beckie, T. Robert, M. Keshavarzi, A. Börner, USDA and ICRISAT for sharing seeds. We thank greenhouse staff at DDPSC and UGA for plant care. We thank Dr. M. Ungerer and two anonymous reviewers for their valuable comment on a earlier version of this manuscript. Most analysis of this work was conducted using the Atmosphere platform of iPlant Collaborative. The iPlant Collaborative is funded by a grant from the NSF (DBI-0735191; Goff *et al.* 2011). This work is funded by DEB-0952185 from the NSF to EAK, DEB-0952177 to KMD, and DE-SC0008769 from the DOE to IB.

References

Baltensperger DD (2002) Progress with proso, pearl and other millets. In: *Trends in New Crops and New Uses* (eds Janick J, Whipkey A), pp. 100–103. *Proc. New Crops and New Uses Strength in Diversity, 5th.*, Atlanta, GA. ASHS Press, Alexandria, Virginia.

Barton L, Newsome SD, Chen F-H *et al.* (2009) Agricultural origins and the isotopic identity of domestication in northern China. *Proceedings of the National Academy of Sciences*, **106**, 5523–5528.

Bennetzen JL, Schmutz J, Wang H *et al.* (2012) Reference genome sequence of the model plant Setaria. *Nature Biotechnology*, **30**, 555–561.

Bomblies K, Yant L, Laitinen RA *et al.* (2010) Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of *Arabidopsis thaliana*. *PLoS Genetics*, **6**, e1000890.

Brachi B, Morris GP, Borevitz JO (2011) Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology*, **12**, 232.

Bradbury PJ, Zhang Z, Kroon DE *et al.* (2007) TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, **23**, 2633–2635.

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, **81**, 1084–1097.

Brutnell TP, Wang L, Swartwood K *et al.* (2010) *Setaria viridis*: a model for C4 photosynthesis. *The Plant Cell*, **22**, 2537–2544.

Caicedo AL, Williamson SH, Hernandez RD *et al.* (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genetics*, **3**, e163.

Cao J, Schneeberger K, Ossowski S *et al.* (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, **43**, 956–963.

Cingolani P, Platts A, Wang LL *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, **6**, 80–92.

D' Ennequin MLT, Panaud O, Toupance B, Sarr A (2000) Assessment of genetic relationships between Setaria italica and its wild relative *S. viridis* using AFLP markers. *Theoretical and Applied Genetics*, **100**, 1061–1066.

Dekker J (2003) The foxtail (Setaria) species-group. *Weed science*, **51**, 641–656.

Doust AN, Devos KM, Gadberry MD, Gale MD, Kellogg EA (2004) Genetic control of branching in foxtail millet. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 9045–9050.

Doust AN, Kellogg EA, Devos KM, Bennetzen JL (2009) Foxtail millet: a sequence-driven grass model system. *Plant Physiology*, **149**, 137–141.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, **6**, e19379.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, **7**, 574–578.

Fournier-Level A, Korte A, Cooper MD *et al.* (2011) A map of local adaptation in *Arabidopsis thaliana*. *Science*, **334**, 86–89.

François O, Blum MGB, Jakobsson M, Rosenberg NA (2008) Demographic history of european populations of *Arabidopsis thaliana*. *PLoS Genetics*, **4**, e1000075.

Goff SA, Vaughn M, McKay S *et al.* (2011) The iPlant collaborative: cyberinfrastructure for plant biology. *Frontiers in Plant Science*, **2**, 34.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, **5**, e1000695.

Hartl DL, Clark AG (2007) *Principles of Population Genetics*. Sinauer Associates, Incorporated, Sunderland, Massachusetts.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology*, **33**, 54–78.

Huang X, Kurata N, Wei X *et al.* (2012a) A map of rice genome variation reveals the origin of cultivated rice. *Nature*, **490**, 497–501.

Huang P, Molina J, Flowers JM *et al.* (2012b) Phylogeography of Asian wild rice, Oryza rufipogon : a genome-wide view. *Molecular Ecology*, **21**, 4593–4604.

Jia G, Huang X, Zhi H *et al.* (2013a) A haplotype map of genomic variations and genome-wide association studies of

agronomic traits in foxtail millet (Setaria italica). *Nature Genetics*, **45**, 957–961.
Jia G, Shi S, Wang C *et al.* (2013b) Molecular diversity and population structure of Chinese green foxtail *Setaria viridis* (L.) Beauv. revealed by microsatellite analysis. *Journal of Experimental Botany*, **64**, 3645–3656.
Jiang H, Barbier H, Brutnell T (2013) Methods for performing crosses in *Setaria viridis*, a new model system for the grasses. *Journal of Visualized Experiments*, **80**, e50527.
Jiao Y, Zhao H, Ren L *et al.* (2012) Genome-wide genetic changes during modern breeding of maize. *Nature Genetics*, **44**, 812–815.
Jørgensen S, Mauricio R (2004) Neutral genetic variation among wild North American populations of the weedy plant *Arabidopsis thaliana* is not geographically structured. *Molecular Ecology*, **13**, 3403–3413.
Kim S, Plagnol V, Hu TT *et al.* (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics*, **39**, 1151–1155.
Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, **9**, 29.
Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.
Li P, Brutnell TP (2011) *Setaria viridis* and *Setaria italica*, model genetic systems for the Panicoid grasses. *Journal of Experimental Botany*, **62**, 3031–3037.
Long Q, Rabanal FA, Meng D *et al.* (2013) Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nature Genetics*, **45**, 884–890.
Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, **11**, 499–511.
McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.
Morton BR, Bi IV, McMullen MD, Gaut BS (2005) Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. *Genetics*, **172**, 569–577.
Nordborg M, Hu TT, Ishino Y *et al.* (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology*, **3**, e196.
Oka H-I, Morishima H (1967) Variations in the breeding systems of a wild rice, Oryza perennis. *Evolution*, **21**, 249–258.
Ossowski S, Schneeberger K, Lucas-Lledo JI *et al.* (2009) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, **327**, 92–94.
Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics*, **2**, e190.
Platt A, Horton M, Huang YS *et al.* (2010) The scale of population structure in *Arabidopsis thaliana*. *PLoS Genetics*, **6**, e1000843.
Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One*, **7**, e32253.
Remington DL, Thornsberry JM, Matsuoka Y *et al.* (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences*, **98**, 11479–11484.
Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, **22**, 939–946.
Romay MC, Millard M, Glaubitz JC *et al.* (2013) Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biology*, **14**, R55.
Sachs MM (2009) Cereal germplasm resources. *Plant Physiology*, **149**, 148–151.
Sage RF (2004) The evolution of C4 photosynthesis. *New Phytologist*, **161**, 341–370.
Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807–820.
United Nations (2013) *World Population Prospects: The 2012 Revision, Press Release*. UN. Available from http://esa.un.org/wpp/documentation/publications.htm.
Van Heerwaarden J, Doebley J, Briggs WH *et al.* (2010) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proceedings of the National Academy of Sciences*, **108**, 1088–1092.
Von Caemmerer S, Quick WP, Furbank RT (2012) The development of C4 rice: current progress and future challenges. *Science*, **336**, 1671–1672.
Wang R, Wendel JF, Dekker JH (1995) Weedy Adaptation in Setaria spp. I. Isozyme analysis of genetic diversity and population genetic structure in *Setaria viridis*. *American Journal of Botany*, **82**, 308–317.
Wang ZM, Devos KM, Liu CJ, Wang RQ, Gale MD (1998) Construction of RFLP-based maps of foxtail millet, *Setaria italica* (L.) P Beauv. *Theoretical and Applied Genetics*, **96**, 31–36.
Wang C, Chen J, Zhi H *et al.* (2010) Population genetics of foxtail millet and its wild ancestor. *BMC Genetics*, **11**, 90.
Wang C, Jia G, Zhi H *et al.* (2012) Genetic diversity and population structure of Chinese foxtail millet *Setaria italica* (L.) Beauv. Landraces. *G3; Genes | Genomes | Genetics*, **2**, 769–777.
Zhang G, Liu X, Quan Z *et al.* (2012) Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. *Nature Biotechnology*, **30**, 549–554.

M.F., M.E., H.Z., X.D., K.M.D. and E.A.K. participated in assembling the *Setaria* collection used in this study; M.F., S.S., I.B., B.A.B. and K.M.D. conducted G.B.S.; P.H. performed the analysis; P.H., K.M.D. and E.A.K. wrote the manuscript.
